# Summary Report (06/30/2014)

## Data Exploring

### Overview

* Week: indicate the time when the survey was conducted
* HasSummary: whether the data has a summary given by the TA
* # of student: the number of students who submit their result
* # of Male: the number of male students
* # of submission
  + Point of Interest (POI): The number of students who submitted the result for POI
  + Muddiest Point (MD): The number of students who submitted the result for MD
  + Learning Point (LD): The number of students who submitted the result for LD
* # of sentences in the TA’s summary
  + POI: the number of sentences in the TA’s summary about POI
  + MD: the number of sentences in the TA’s summary about MD
  + LD: the number of sentences in the TA’s summary about LD
* Average: the average number of students who submitted the results / # of sentences for the weeks that have TA’s summaries
* Std: the standard deviation of number of students who submitted the results / # of sentences for the weeks that have TA’s summaries
* Correlation: the correlation between the number of students who submitted the results for the weeks and the number of sentences in the TA’s summaries

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | # of submission | | | # of sentences in the Summary | | |
| Week | Has summary | # of student | # of Male | Point of Interest | Muddiest Point | Learning Point | POI | MD | LD |
| 1 | True | 53 | 34 | 42 | 42 | 42 | 3 | 3 | 3 |
| 2 | True | 49 | 32 | 45 | 45 | 45 | 3 | 4 | 2 |
| 3 | True | 50 | 30 | 45 | 45 | 45 | 4 | 4 | 4 |
| 4 | True | 50 | 31 | 45 | 45 | 45 | 2 | 3 | 3 |
| 5 | True | 50 | 31 | 42 | 42 | 42 | 3 | 3 | 3 |
| 6 | True | 50 | 31 | 43 | 42 | 42 | 2 | 4 | 4 |
| 7 | True | 51 | 31 | 44 | 44 | 44 | 3 | 4 | 3 |
| 8 | True | 50 | 32 | 42 | 42 | 42 | 4 | 3 | 3 |
| 9 | True | 50 | 32 | 37 | 37 | 37 | 8 | 12 | 3 |
| 10 | True | 53 | 33 | 36 | 36 | 36 | 5 | 6 | 3 |
| 11 | True | 50 | 31 | 36 | 36 | 36 | 9 | 8 | 9 |
| 12 | True | 50 | 31 | 43 | 43 | 43 | 4 | 5 | 3 |
| 13 | False | 50 | 31 | 36 | 36 | 36 | 0 | 0 | 0 |
| 14 | False | 51 | 31 | 35 | 35 | 35 | 0 | 0 | 0 |
| 15 | False | 49 | 31 | 33 | 33 | 33 | 0 | 0 | 0 |
| 16 | False | 50 | 31 | 35 | 35 | 35 | 0 | 0 | 0 |
| 17 | False | 49 | 31 | 37 | 37 | 37 | 0 | 0 | 0 |
| 18 | False | 49 | 31 | 23 | 23 | 23 | 0 | 0 | 0 |
| 19 | False | 52 | 31 | 36 | 36 | 36 | 0 | 0 | 0 |
| 20 | False | 50 | 32 | 35 | 35 | 35 | 0 | 0 | 0 |
| 21 | False | 50 | 32 | 37 | 37 | 37 | 0 | 0 | 0 |
| 22 | False | 49 | 31 | 34 | 33 | 33 | 0 | 0 | 0 |
| 23 | False | 47 | 28 | 19 | 19 | 19 | 0 | 0 | 0 |
| 24 | False | 50 | 30 | 35 | 35 | 35 | 0 | 0 | 0 |
| 25 | False | 50 | 32 | 35 | 35 | 35 | 0 | 0 | 0 |
| Average |  |  |  | 41.7 | 41.6 | 41.6 | 4.2 | 4.9 | 3.6 |
| Std |  |  |  | 3.4 | 3.4 | 3.4 | 2.2 | 2.7 | 1.8 |
| Correlation |  |  |  |  |  |  | -0.8 | -0.7 | -0.5 |

## Possible Approaches/Contributions

### Clustering

One key challenge of this data is that the number of sentences in the summary changes. This is different from traditional summary which has consistent summary length/compression ratio. However, the number of sentences in the summary has no correlation between the length of input documents (The results submitted by the student).

To address this issue, we can take advantage of clustering algorithms like KMeans, which can be set to different number of clusters according to a certain rule.

Another way to address this issue is to use the Hierarchical Summarization proposed by [1]. However, the proposed hierarchical summarization used time information to cluster documents and we might need to cluster them by similarity or topics.

### Extraction

This is the most commonly used type of summarization. We can use try the MEAD summarization toolkit first [3]. And then, we can try similar techniques to proposed by [4] because Muddiest Point Coding is similar to the helpfulness score. In advance, we can try the state-of-the-art technique by approaching summarization as an ILP problem [2].

### Keywords

Another way to give a summary is to visualize the keywords in the documents using WordCloud [5] or topicCloud [6].

# Reference

1. Soderland, Janara Christensen Stephen, Gagan Bansal Mausam, and India Delhi. "Hierarchical Summarization: Scaling Up Multi-Document Summarization." ACl 2014
2. Thadani, Kapil, and Kathleen McKeown. "Sentence compression with joint structural inference." In *Proceedings of CoNLL*. 2013.
3. [www.summarization.com/mead](http://www.summarization.com/mead)
4. Xiong, W. (2013). Helpfulness-guided review summarization. Proceedings of NAACL-HLT Student Research Workshop, pages 77–83. Atlanta, Georgia.
5. http://www.wordle.net/
6. Liu, Shixia, et al. "Tiara: Interactive, topic-based visual text summarization and analysis." ACM Transactions on Intelligent Systems and Technology (TIST) 3.2 (2012): 25.